数理统计学:世纪末的回顾与展望

陈希孺

ABSTRACT

In this article ,the author summarizes the progress of Mathematical Statistics in this century and the orientation in 21st century. The author concludes that the movement of the subject is similar to which in the beginning of 20th century ,while there are also substantial difference.

关键词: 数理统计; 回顾; 展望

一、20 世纪数理统计学发展概述

20 世纪,特别是其上半叶,是数理统计学发展史上一个辉煌的时代。从现代数理统计学框架的建立到发展为一个成熟的学科,是在这个时期完成的。20 世纪初,数理统计学面临一个转折点,意思是它必须有新的突破才能获得进一步发展的契机。20 世纪早期一批以费歇尔为首的统计学大师成功地应对了这个局面,创造了非凡的业绩。按照国际上一些知名统计学家的看法,20 世纪末数理统计学发展的态势,与世纪初颇有相似的地方。人们在呼唤"21 世纪的费歇尔"。当然,广义地说,这也是每一位数理统计工作者所肩负的任务。中国作为一个世界大国,年轻一代的数理统计学者应该也有条件在这方面作出自己的贡献。

为了更清楚阐述上文的意思,需要对数理统计学的历史作一个简短的回顾。按目前数理统计学界公认的看法,数理统计学是"收集和分析带随机性的数据的科学和艺术"。以笔者的看法,这个内涵规定了它是一个中立性的工具。"中立"的意思是指这门学科不带任何社会的、政治的或意识形态上的倾向性,因而也不存在它自成学派或从属于何学派的问题。有一种看法认为社会经济统计学与数理统计学是"大统计学"中的两个对立的学派。笔者认为这种看法值得商榷。的确,在社会经济统计学中该不该使用数理统计方法,在哪些问题上或者在何种程度上应否使用数理统计方法,是可能存在不同意见的。如果说由于对这些问题的看法不同而有学派存在,那还算言之成理。但这些问题与数理统计学无关:数理统计学只是一种工具,谁如觉得这个工具对他有用,就可以使用它——当然在使用中必须遵守这门学科的规范,否则就可能产生误导公众及提供错误的决策依据的后果。历史上(部分地直到如今)数理统计方法曾遭到一些批评和怀疑,一定程度上与上述情况有关。

数理统计学起源于何时?这是一个无法也不必做出定论的问题。有的学者把英国学者格朗特的著作《关于死亡公报的自然和政治观察》发表的年份 1662 年定为这门学科的诞生之日,恐怕也只能算是一家之见。实际情况是,可以说直到 20 世纪初,并不存在一门统一的数理统计学科,而只是在各实用领

域中的学者因工作上的需要而分头发展了一些分析数据的方法,即统计方法。最主要的有3个方面:一是天文和测地学中因误差分析问题而导致最小二乘法和正态误差的发明。起初,人们认为"误差分析"与"统计分析"是根本不同的两回事:前者的数据是对一个对象多次测量所得;后者的数据则是对多个对象各测量一次所得。按现今的数理统计学框架,我们容易认识这是一回事,但在当时则不然。到19世纪中、后期,经过凯特勒、盖尔顿等在社会学和生物学方面的实际工作,以及埃其渥斯、卡尔·皮尔逊等的数学理论工作,终于把二者统一起来,并在20世纪得到发扬光大。直至如今,线性模型———最小二乘法——正态误差这个体系下所发展的方法,在相当大的程度上仍占据了应用统计方法中的主导地位。所以有人说,天文学是数理统计学的母亲。

第二个方面是人口学。前文提到的格朗特的著作是一个重要例子。这个方向发展了离散数据统计,即以二项分布和波哇松分布为代表的统计方法。另一个重要之点是它在19世纪即开始孕育了抽样调查的思想。这也在20世纪得到发扬光大,成为现今统计方法中的重要组成部分。有的统计史学家评说:19世纪的统计就是频率分析。那是因为,当时处理误差分析的一套工具尚未被视为属于统计方法的范畴。

最后一个方面是生物学,特别是遗传学。英国学者盖尔顿在1874到1890年间的工作,引进了相关和回归的思想。其重大意义在于它开创了分析多维数据的统计方法。此前的统计方法都是单指标性的,不能顾及指标间的相互关系。而在实用问题中一般涉及多个彼此相依的指标,孤立地分析单个指标无法得出符合实际的结论。盖尔顿的工作经过埃其渥斯、卡尔·皮尔逊和约尔在数学上的整理,到20世纪又经过费歇尔等一批学者的深化,直到目前仍不失为应用统计方法中的重镇和理论统计学中的主流方向之一。

有人把上面粗略描述的,大体上到19世纪末为止的统计学的发展图景作了一个小结,归纳为以下3点:(1)统计方法是基于实用的需要,在不同领域中分头发展的。(2)没有专职的(以统计学为主业的)统计学家。对统计方法作出重大贡献的人,其主要身份是某个其他领域的学者,这在公认是现代数理统计学的奠基者费歇尔和卡尔·皮尔逊身上还可以看出来。(3)统计学没有一个严整的学科框架。费歇尔传记的作者J. F. Box在谈到20世纪初期统计学状况时曾提到,当时在人们的意识上连参数与统计量都没有严格区分开。有的学者提到,当时在统计方法的工具袋里已有了一些积累,包括最小二乘法(平均值可视为其特例)、方差、频率、二项分布、误差理论和正态分布、相关回归、矩估计、皮尔逊曲线族以及稍后的Student t分布等。但它们是一些不连贯的片段,缺乏一个完整体系。

所以,在20世纪初年,摆在数理统计学面前的重大问题是建立一个理论(数学)上的框架。它不仅能包容已有的成果,而且还要对未来努力的方向起指引的作用。如大家所知道的,这个任务由以费歇尔为代表的一班统计学大师出色地完成了。这些统计学大师中除费歇尔外,还可以算上爱根·皮尔逊、奈曼及较晚的瓦尔德。至于卡尔·皮尔逊,有一种看法认为它是"旧统计"的押阵大将。但平心而论,他的工作,尤其是1900年发表的关于拟合优度检验的论文,对"新统计"的诞生有着不可低估的影响具有划时代的意义。至于费歇尔,其贡献更是全方位的:在理论方面,他分别于1921年和1925年发表的论文《理论统计学的数学基础》和《点估计理论》,奠定了统计学的大体上沿用至今的数学框架;在方法的层面,他提出的似然估计、试验设计与方差分析以及一大批小样本抽样分布的结果,迄今仍有着重大的影响。其业绩在

20 世纪统计界确实无人可比。所以美国统计学家埃夫龙在 1996 年一篇论文中把他比作“统计学的凯撒”。

前文提到,临近 20 世纪末,数理统计学发展的态势,颇有与世纪初相似之处。这一点要联系到 20 世纪下半叶数理统计学的发展状况来讨论。

1940 年,以克拉美的《统计学的数学方法》一书的出版为标志,数理统计学被公认为已形成一门严整的数学学科——应当注意的是:这一点固然与费歇尔等人为统计学制定了合适的数学框架有关,更本质的原因在于统计学中的“数据”已超脱了其实际含义:一组数据如假定来自正态总体,则与此有关的方法(如 t 区间估计、F 检验等)都可以使用,而无须顾及数据从何而来。正如在数学中人们说 1 + 2 = 3,而不必顾及这 1、2、3 是什么一样。

数理统计学一经数学化,就有其自身的发展规律,一般认为,一个数学分支中新问题的来源,有“外生”和“内生”两种。前者是因外部的需要,一般是实际应用中的需要所提出的问题,而后者则是由学科的“自我扩张”引起的问题,不必有其实际背景。如前所说,在较早的时期(约在 20 世纪 30 年代或放宽一些到 50 年代),数理统计学与实用紧密结合,所研究的问题以“外生”性的为主。此后,情况有了很大变化:相当大部分的统计学理论研究转向“内生”性的问题,以“在预设的模型下寻求符合某种准则的最优解”及“大样本理论”两个方向为代表。应当指出的是:并非说沿着这些方向所作的工作全无实际意义。有些工作(主要在较早时期)是以往比较粗糙的结果的完善。例如有关极大似然估计的渐近性质,费歇尔在 1925 年关于点估计的论文中就有初步的讨论。到五、六十年代,在数学上得到更完满的发展。这类工作兼有理论和实用两方面的意义。有的在优化理论框架下得出的结果,如算术平均值或更一般地最小二乘估计在种种条件下的优良性质的结果,虽则对应用统计方法无所增添,但深化了我们对这些重要方法性质的了解,也是很有意义的。至于大样本理论,其大量的繁琐结果可说已趋于末流——既无理论上的数学美,又对分析数据不起什么作用。但也不可否认,其中也颇有些富有实际意义的结果,特别是非参数统计有关的一些大样本结果,为在免除正态假定下进行数据分析提供了可用的替代方法。

虽然可举出以上这些有利情况,但不能不承认,从总体上说,由这些“内生”问题产生的结果,多数是与数据分析没多大关系,从纯数学的角度看也缺乏深度。这种情况引起了不少统计学家的忧虑和反思,以至有所谓“统计学危机”的呼声。

以上的简略描述表明,数理统计学在 20 世纪下半叶,理论上缺乏有意义的、突破性的进展。实用的或方法层面上的情况如何?应该说有不小的成绩。其中一部分得力于功能强大的计算机,它使一些需要大规模计算的方法能付诸实用,从而大大拓展了统计方法的应用面。在方法本身的研究上也有不少进展。不久前出版的一本论述“统计学中的突破”的著作,列举了到 1980 年为止统计学方面的 40 项“突破”,就其内容看(如赤池弘次的 AIC 准则,维尔考克森的秩和检验之类)大都是局部范围内的方法性的成果,并非有全局意义的“突破”。统计学家休伯 1997 年在北京的一次讲演,认为近几十年来数理统计学只有 3 项值得一提的重要成果:其一是他自己发展的稳健统计(这概念可追溯到费歇尔在 1920 年的一项关于比较绝对平均差和标准差的优劣的工作),另有埃夫龙在 1979 年提出的“自助法”(bootstrap)和生存分析。若情况果真如此,则 20 世纪下半叶统计学的成绩可说是很暗淡了。依笔者所见,情况要乐观一些,比如回归分析和多元分析中诸多的理论和方法进展、模型选择、试验设

计、生存分析、贝叶斯统计等方面,都颇有一些富有实用意义的成果。

但不容否认的是,20 世纪下半叶数理统计学方面的成就,主要限于若干局部性的、具体问题的方法性的层面上,全局性的、涉及根本的统计思想的成果,绝无仅有,拿一句人文科学讨论中常提到的套话来形容,可说是"学问家凸显,思想家淡出"。

以上种种情况使不少统计学家认为,统计学又面临一个新的突破的形势,或者也可以说,到了一个需要变革的时期,这与 20 世纪初的情况有其相似之处。

二、数理统计学未来的发展

这种突破会指向何方?要采取怎样的措施以有利于促成这种突破或变革?自 20 世纪 60 年代以来,不少学者,通过在有关会议上发表讲演或在刊物上发表论文,表达了各自的看法。有些看法有很大的一致性,例如主张统计学要回到以前那种重视联系实际的传统;主张"推倒围墙",即重视与其他学科的交流和渗透;主张在统计教育上实行与此相应的变革等。在预测未来发展的主流上,则多有分歧。下面对一些较有影响的观点择要介绍一下。

1. 数据分析。美国资深统计学家图基在 1962 年发表了一篇题为《数据分析的未来》的长文,大约"数据分析"一词即起源于此文。这是第一次由一个极有影响的统计学家对当时的数理统计学发展状况作出反思并提出一种变革的方向,因此有重要的意义。直到今天,该文所表述的观点还经常得到统计学家们的征引。

此文主要的精神是对当时统计界流行的以模型为出发点的做法提出反思,主张让数据多起作用。模型应当从分析数据中产生而不应让数据去曲合预设的模型。为此他主张对现行统计学的根基"用概率刻画统计推断的不确定性"作出松动——概率只是作为一种工具而非基础,适合使用时则用,不合用时就不用。提出这一主张与打破"以模型为出发点"或"预设模型"有关:没有一个充分简化的模型,就无法对统计推断的概率性质作深入探讨,以致一些"为发表文章而作研究"的工作不能不预设模型。除此以外图基还提出了一些原则性的主张。如要研究有现实意义的新问题,在更现实的条件下研究老问题,把统计学定位为一门科学而非数学——这意味着实用性优先于推理的严格性等等。图基及其支持者以后在一些文章和专著中进行一步阐发了他们的主张。所有这些人们现在将其归在尚未成形的"数据分析"的名目下。

笔者认为,虽然目前讲图基的主张在统计学界的支持率还不能算是很高,但由于以下两个情况,在未来可能发展为一个很有影响的思潮,并在相当程度上改变现行统计学的面貌。一是功能强大的计算机的广泛使用;二是在各个领域里不断提出的带有复杂结构数据的问题,如高维模拟仿真、模式识别、图像和信号处理、人工智能、神经网络、数据采掘(data mining)等。由于数据的随机性,这类问题在一定程度上和统计学有关,又因其复杂性使概率方法难于充分有效地使用,因此可能需要某种"折衷"的办法:既考虑到数据的大量性和复杂性而不能拘泥于一定的概率模式,又能使因数据的随机性而产生结论的不确定性有着某种科学的评价标准。也许这是一个使现行统计学产生"突破"的地方?

2."边缘学科"。如前所说,在 19 世纪末之前,统计学尚未成为一个今天意义下的独立学科,其发展是为应付现实的需要,结合其他学科来进行的。近若干年来,这个发展模式受到一些有影响的统计学家的推崇,认为有可能是将来的主流模式。统计学家休伯 1997 年在北京的一次讲演中,把统计学发展的历程画成一条螺旋线而非直线,意谓其发展不是直线式的,而是可能具有某种"回归"

的性质或我们常说的"螺旋式进展"——当然是在提高的基础上回归而非简单重复。他还发表了一个"盛世危言"式的见解:如果统计学家的研究成果脱离实际应用的状况得不到改变,则这种形式的统计学将走向消亡,人们将像以往那样回到各个学科去发展适用的统计方法。这类方法不必具有通用的性质(比如像回归分析、方差分析这类统计方法,都具有很广的通用性)。

他描述的这种图景眼下大概还不会成为现实——相信"通用的"统计方法仍有很大的发展余地,数理统计学作为一门独立学科的地位还没有动摇的迹象,但其思想则颇有可取之处。"通用方法"的发展也不可能是纯数学思维、闭门造车式的。如费尔顿——皮尔逊发明的相关回归,费歇尔发明的方差分析这类"通用"方法,是结合像遗传学和农业试验的需要而得到。一位有名的华人生物统计学家曾指出搞统计必须结合一个 area,也是这个意思。另一位美籍华人学者李景均(C. C. Li)教授因研究群体遗传学的需要而发明被称为"路径分析"(path analysis)的统计方法,曾应某刊物之约发表长达 80 余页的专题论文,成为该领域国际公认的权威。其方法不止适用于遗传学。他并不以数学见长,如果投身到统计学的纯理论研究,也许不一定能做出达到这个水平的成就。总之,历史和现实都证明了:统计学和其他学科结合发展是一个正确的方向,也极可能成为未来发展的主流之一。

3. 贝叶斯统计。频率学派和贝叶斯学派的对立是 20 世纪数理统计学发展中一道亮丽的风景线。临到世纪末,早期那种情绪性的对立局面似乎已逐渐消退。原因之一是早期那些大师都已去世或淡出舞台,后继者不一定那么执着于"纯哲学"式的争论。也因为经过几十年的实践,统计学界大体上有了一种共识:至少在参数统计的范围内,这两个学派所达致的结论,所提供的方法基本相似。另外在局部范围而言,两派的方法也确实各有短长。

英国老牌的贝叶斯派统计学家在近年的一次访谈中预言,21 世纪将是贝叶斯统计一统天下的局面。在另一个场合中他把时间具体化到 2020 年。这后一点看来不像会成为现实。但近年来统计刊物上发表的一些学者的见解,确给笔者这样一个印象,即贝叶斯学派正在取得上风。这有多方面的原因,不在此细论。其中一个因素可能是:在对 20 世纪统计学的状况进行反思时,几乎所有的负面因素都与频率学派有牵涉,如过分的数学化而形成"两张皮"的现象。其中尤以将统计问题归结为最优化数学问题的见解倍受非议,有人讥之为"错误问题的正确解答。"

如果贝叶斯统计真成为主流,在未来世纪它的主攻方向如何?有的学者也对此发表了见解。

如所周知,贝叶斯学派有"主观"和"客观"两个系统。主观贝叶斯学派认为先验分布的选择纯是使用者个人的事,不可能也不应该去寻求某种公认的、"客观"的选择。进一步的引伸是统计推断纯粹是主观行为,不可能用一种科学标准去规范它。学者们认为,这一学派会有其存在余地。它主要适用于经济决策中。在这种问题中,决策主体的条件和掌握信息资源的不同当然会影响其做出的决策,不可能有统一的标准。但是在科学研究(以寻求客观真理为目标)性质的问题中,或一般地说,在主观因素影响较少的问题中,这种思维模式恐难于为人们所接受。

客观贝叶斯学派主张的核心在于给先验分布的选定制定一种大家都遵守的"客观"规则,而不由人随意地主观选定。这里所谓"客观",不应理解为在频率意义下与实际情况符合——即同类问题大量出现而按参数值的频率去确定先验分布(果真有这个情况,问题可纳入频率学派的体系下)。因为在绝大多数情况下,问题是一次性的,不存在按这种方式

决定先验分布的可能。因此,真正的贝叶斯派所持的立场是:先验分布的选取纯粹因为它是统计推断工作得以完成的一个必须的成分,不存在它与现实符合与否的"客观性"问题。

客观贝叶斯学派源起于贝叶斯本人,后曾被拉普拉斯广泛使用,所以也有人把它称为拉普拉斯学派(相应地,主观贝叶斯学派有时被称为芬内迪—赛瓦奇学派)。其时公认的先验分布是根据"同等无知"原则而确定的均匀分布。后来的学者用稍广一些的"无信息先验分布"来取代它。费歇尔早就指出过这样确定先验分布的一个问题,即随着参数取法的不同会导致不同的先验分布。针对这一点杰弗里斯引进了一种选择先验分布的方法,可以避免这个困难,但仍不是能很令人满意。例如对二项分布 $B(n,\theta)$ 的参数 θ $(0 \leq \theta \leq 1)$,按"同等无知"原则取先验密度 $p(\theta) = 1$,而按杰弗里斯的方法则应为 $p(\theta) = (\theta(1-\theta))^{-1/2}$。从实用的角度看,这个选择似乎不比 $p(\theta) = 1$ 更有吸引力。

有些学者认为,21 世纪统计学的一个动向是"频率学派与客观贝叶斯学派的合流"。而在这个进程中,有两个问题是主要的:一是研究更合理的制定先验分布的准则。这里"合理"的含义恐怕主要不是某种抽象的理论标准,而是看在这个体制下所作出的统计推断的合理性即在应用上的有效性。在此,已有的大批"经过考验的"成果会成为一个重要的参照标准。另一个问题是在非参数模型之下,亦即在函数空间中如何确定先验分布。早在 20 世纪 70 年代有人作过尝试,引进狄利希莱分布去定义非参数先验分布,后来也没有值得注意的进一步发展。这个问题无疑是一个富有挑战性的困难问题。

从广义的意义说,费歇尔在 1930 年提出的"信任推断"可纳入贝叶斯学派的体系内。费歇尔一贯不赞成先验分布的提法,但却接受了贝叶斯学派的核心思想——由样本产生一个关于参数的分布,这在贝叶斯学派中称为后验分布。费歇尔提出"信任推断"的客观效果等同于一种不要先验分布的贝叶斯统计。由于一些内在的困难,几十年以来沿着这个方向没有取得多少进展,不少统计学家把费歇尔的主张看作他诸多成就之下的一个引人注目的失败。不过仍有少数学者,如费雷塞,坚持在这个领域工作。最近关于这个问题的兴趣有复活的趋势。有人甚至预言,费歇尔在 20 世纪提出的这个倍受冷落的概念,很可能在 21 世纪开花结果。

以上所论的是一些属于大的趋势方面的问题。至于更具体领域中可能的进展,也有学者讨论。如模型选择,基于似然函数和条件推断的结构;离散和非线性多元分析;不完全数据分析;(广义的)经验贝叶斯方法;定性、不回答和缺落数据的处理等。因篇幅关系均不一一细论。笔者总的看法是:虽则 20 世纪末统计学发展的态势确有某种与世纪初相似的地方,但仍有着根本的不同。这种不同表现在世纪初时数理统计学的学科框架尚未建立,而目前已在相对成熟阶段上走了相当一段距离。局部性的重要成就时有发生,而全局性的甚至根本改变本学科面貌的那种突破,在可预见的将来不大可能发生:"21 世纪的费歇尔"产生的时机还未到来。

作者简介:陈希孺,男,1934 年出生,1956 年毕业于武汉大学数学系,获理学学士学位。同年到中国科学院数学所工作。1957—1958 年间在波兰科学院学习数学。1960—1986 年在中国科技大学从事教学工作。1986 至今在中国科技大学研究生院从事教学和科研工作。1986—1988 年在美国匹兹堡大学作访问学者,1992 年在加拿大约克大学作访问学者。现任中国科技大学教授、中国科学院院士、中国统计学会副会长、中国现场统计学会理士长。主要研究方向为大样本理论。

(责任编辑:许亦频)